ChIP-seqデータベースの構築による遺伝子転写制御機構の解明

氏名	安澤（あんざわ）　隼人（はやと）
学位の種類	博　士（情報科学）
学位記番号	情　博　第　７１９号
学位授与年月日	令和　２年　３月２５日
学位授与の要件	学位規則第４条第１項該当
研究科、専攻	東北大学大学院情報科学研究科（博士課程）応用情報科学専攻
学位論文題目	ChIP-seq データベースの構築による遺伝子転写制御機構の解明
論文審査委員	（主査）東北大学教　授　木下　賢吾 東北大学教　授　田中　和之　　東北大学教　授　中尾　光之

論 文 内 容 の 要 旨

第Ⅰ部 緒論

生命現象において主たる遺伝情報である DNA の塩基配列が共通であるにも関わらず、多様な表現型を顕現しうる例が多く知られている。その原因の 1 つとして、塩基配列が同一でも DNA 結合タンパクやその化学的修飾による遺伝子発現の制御が挙げられる。このような転写制御機構の解明は、生命現象を理解する上で重要な課題であるといえる。

ChIP シーケンシング（ChIP-seq）法は DNA 結合タンパクの結合領域をゲノムワイドに検出する手法として、エピジェネティクス分野を始めとした生命科学分野に膨大な知見をもたらしており、現在もそのデータは増加し続けている。ChIP-seq 法はクロマチン免疫沈降法（ChIP: Chromatin ImmunoPrecipitation）と次世代シーケンサー（NGS）による塩基配列の超並列解読技術を応用したものであり、従来手法と比較して革新的なデータのハイスループット化とコストの低減を実現した。一方で、ChIP-seq は RNA-seq のような他の NGS アプリケーションと比較すると、ChIP の過程に起因する実験的複雑さからシグナル-ノイズ比（SN 比）が低くなりがちであり、クオリティーコントロール（QC）の重要性は高い。

転写制御は多くの場合複数の転写制御因子の協調により達成されることが多く、ChIP-seq 実験から制御機構の全体像を明らかにするためには、複数の実験結果を統合することが求められている。こうした背景を元に、先行研究で公開済みの公共 ChIP-seq データを利用した ChIP-seq データベースが構築されるようになった。しかし、既存のデータベースは、NGS の生データを格納した一次データベースの整備状況や、先述した QC の必要性を背景に、網羅性・信頼性を十分に両立しているとは言えない。本博士論文は、この問題点に対処し ChIP-seq データドリブンに生物学的な知見や研究仮説を提供するプラットフォームとなる ChIP-seq データベースの構築を目標として研究を行った。

第Ⅱ部 ChIP-seq データ解析パイプラインの開発

ChIP-seq データを解析するためのパイプライン（ワークフロー）は既に基本的なアプローチが確立されているものの、公共データに基づくデータベースを構築するためにはいくつかの問題が生じる。1 つはメタデータ（付帯情報）の不完全性であり、規格化されていない自然言語に近いメタデータは自動処理の難易度を上げ、網羅性を下げる要因であった。もう 1 つは QC の担保である。公開されているデータは必ずしもその質が高いとは限らず、再利用に際しては QC の情報も共に提供することが肝要である。

本研究では、前者の問題に対処するために GEO のメタデータについて、特に重要なメタデータであるサンプル名（細胞種・組織）と ChIP ターゲットを検出する手法を開発した。本手法により、ChIP-seq 実験の約 84%に対してこれらの情報を抽出することができた。また QC について、複数のプロジェクトや先行研究を基に指標と基準を設定し、必要な処理をパイプラインに盛り込んだ。特に、Strand cross-correlation を用いた新規指標については第Ⅲ部およびⅣ部にて詳

説する。

また、バリアントコール解析で実績のあったデコイ配列入りリファレンスゲノムについて、ChIP-seq 解析における効果を検証し、難読領域に対する改善効果が認められたため、本解析パイプラインではこれを採用した。

第Ⅲ部 Strand cross-correlation の理論的特性評価

ChIP-seq 実験に対する QC 指標は種々の要因に対する様々な指標が提案されてきたが、特に実験に対する SN 比の評価は ChIP の成否や検出しうるピーク数に対する見積もりとなることから、ChIP-seq 実験の QC において代表的な指標であるといえる。SN 比を反映する指標として一般的に用いられているのは FRiP（Fraction Reads in Peaks）であるが、FRiP には、ピークが既知である必要がある・ピークコールの手法に依存する・取得リード数に依存する、といった欠点がある。これらの問題を克服する指標として Strand cross-correlation を用いた指標が提案され、ENCODE Project のような大型プロジェクト等で実用されてきた。しかし、Strand cross-correlation を基にした手法は観測事実とは合致するものの、指標や基準値の設定は経験則に基づいており、このような指標が明確に何を反映するのか、どのようなデータ間で比較可能なのかという点において不明瞭なままであった。そこで本研究では、ChIP-seq リードの分布のモデル化を通して Strand cross-correlation の理論値を導出し、Strand cross-correlation を用いた QC 手法の特性評価を行った。

ChIP-seq リード分布は次のようなモデルを仮定した。まず、単一の染色体からなる長さG塩基のゲノム中にn箇所の互いに十分距離が離れた結合部位があるとし、これらの両端には幅d塩基を隔てて長さw塩基の順鎖・逆鎖側のリードエンリッチ領域を設定した。次にこのゲノムに対し ChIP-seq 実験を行いM本のリードを取得したと仮定し、またリードのうちちょうど半分ずつがそれぞれ順鎖・逆鎖側にマップされるとした。ここで、SN 比を表現するパラメーターαを導入し、全リードのうち$M\alpha$本がエンリッチ領域にマップされると仮定した。

相互相関関数は先行研究で用いられた NCC（Naïve Cross-Correlation）を使用した。これは、リードの分布を表現する関数にバイナリ関数を用いたものである。本研究では、NCC の取り得る最小値と最大値について、エンリッチ領域に対してリードを過剰に取得した場合（飽和条件）とそうでない場合（不飽和条件）について導出を行った。また、より精確なプロファイリングが可能な手法として同先行研究で提案された MSCC（Mappability-Sensitive Cross-Correlation）についても検討を行い、理想的な条件下では NCC と MSCC はよく一致することが予想された。

導出結果に対して検証するため、シミュレーションデータを用いた検証と実データを用いた検証を行った。また、これらの検証を大規模かつ効率良く実施するため、NCC および MSCC を計算するソフトウェア PyMaSC を新規に開発した。まず、提案したモデルに基づきシミュレーションデータを作成し、モデルから予想される最小値・最大値と実際に計算したそれらを比較した。結果、最大値については不飽和条件・飽和条件下どちらもよく一致した。最小値については NCC では理論値との差異が見られたものの、MSCC ではよく一致したことから、最小値に対する誤差はゲノムの Mappability に起因するものであり、これを補正できる MSCC の方がより検出感度が高いことを発見した。次に、ENCODE Project で公開されている実 ChIP-seq データを用いた検証を行った。ここでは、ピークコールを行うことでn, w, αを推定し、これらから予測した最大値と実際に計算して得られた NCC の最大値を比較した。結果、転写因子をターゲットとした実験は非常によく理論値と実測値が一致し、他の種類の実験についても、強い正の相関が見られた。この誤差はパラメーター推定時に発生した誤差であると考えられ、本研究で用いたモデルが必要十分な表現力を備えていることが示唆された。また、MSCC についても検討した結果、ほとんどの場合で相関係数は NCC とよく一致し予想が確かめられた。

最終的に、NCC の最小値はほとんど 0 になることが予想され、シミュレーションの結果はこれを裏付けた。一方最大値については不飽和条件・飽和条件それぞれ異なるが、実際のデータのほとんどは不飽和条件を満たすことが分かり、NCC を用いたプロファイリングにおいて重要な指標となる最大値は次式で表現できることが明らかになった。

$$\mathrm{Max(NCC)} = \frac{M\alpha^2}{2nw}$$

これらの結果から、Strand cross-correlation を用いた解析では ChIP-seq 解析における重要な指標であるnとαを独立して予測するのが困難である可能性、また従来指標の計算方法が不適切である可能性を示唆した。さらに、この関係性が明らかになったことで NCC の最大値を比較できる条件を明確にすることができた。

第Ⅳ部 Strand cross-correlation を用いた新規品質評価手法の提案と検証

第Ⅲ部で得られた結果から、Strand cross-correlation を用いた新規 QC 指標である VSN（Virtual S/N）を提案した。これは NCC の最大値から既知であるMと NCC の分布から推定可能であると考えられるwを除いたものである。

$$\mathrm{VSN} = \frac{w}{M}\mathrm{Max(NCC)} = \frac{\alpha^2}{n}$$

本来、nとαは独立であるが、実際に得られる解析結果ではよく比例することが確かめられた。これは、実際に検出可能なnは SN 比が良いほど多くなるという自然な関係を反映していると考えられるため、α/nは一定の値を取ることが期待できる。従って、VSN の大小とαの大小が一致するであろうことを期待して VSN を評価指標として提案した。

VSN が評価指標として機能することを確かめるため、まずwを MSCC の計算結果から得られることを確かめた。次に、VSN と FRiP に整合性があるかを、既存の Strand cross-correlation を用いた指標とスピアマンの順位相関係数を用いて比較した。結果、ヒストン修飾をターゲットにした実験ではやや劣るものの、転写因子の場合、あるいはこれらのターゲットが混在する場合には、VSN は既存の指標と比較してより高い整合性を示した。また、Strand cross-correlation を用いた指標の特徴である、総リード数Mに対するロバスト性についても検証を行い、結果 VSN が既存指標と同程度のロバスト性を備えていることを明らかにした。

以上の検証から、VSN が実用的な QC 指標であることが確かめられ、本研究で開発したデータベースでは VSN を QC 指標の 1 つとして採用した。

第Ⅴ部 ChIP-seq データベースの開発と転写制御解析

これらの成果を踏まえて本研究で提案した解析パイプラインを実装し、実際に大規模解析を行った上で得られた解析結果を可視化するデータベース C4S-DB（Comprehensive Collection and Comparison for ChIP-Seq）を Web アプリケーションとして開発した。

ドライ解析における解析手法の再現性・可搬性の問題に対処するため、解析パイプラインの各ステップは Docker を用いてコンテナ仮想化を行った。また本研究ではこのコンテナを用いて解析をクラウドコンピューティング環境である AWS 上で実施した。仮想化技術に基づく並列化により大規模データを高速に処理することができ、またコストの面でも、同規模の解析を行うためにオンプレミス環境を整備し運用するコストと比較すれば十分実用的であることが分かった。

データベースの機能として、個々の実験に対する解析結果の可視化・遺伝子周辺領域における解析結果の可視化・実験データ間の大域的な類似度の可視化の 3 点を軸に実装した。デモンストレーションとして、ENCODE Project から得たヒト A549 細胞 ChIP-seq 実験のデータ 283 件を用いた転写制御解析を行った。まず実験間の類似度をクラスタリングした類似度マップからは、転写活性・インシュレーター形成・デキサメタゾン（DEX）誘導に関するクラスター群が得られた。特に DEX 誘導群について遺伝子周辺領域を確認すると、転写開始点を中心に種々の転写制御因子が共結合している様子を確認できた。一方、これらの結合状況から制御関係をネットワークとして図示した結果からは、DEX 受容に始まる明確なカスケードを確認するには至らなかった。

したがって、現段階で実装した機能により個々の解析結果やゲノム領域の可視化や転写制御因子間の関係性を観察する最低限の機能は提供できるものの、より本質的な解明に迫るためには実験条件に基づいた分析や文献情報等に基づくアノテーションとの統合など今後さらなる改善が必要であろう。

第Ⅵ部 総括

本論文では、公共 ChIP-seq データを利用した転写制御機構の解明に繋がる ChIP-seq データ

を開発・提案した。本研究で開発した解析パイプラインは、既存のデータベースに欠けていた網羅性・信頼性を改善してより大規模な ChIP-seq データセットの作成を可能にするものである。また、解析結果を可視化する Web アプリケーションとして C4S DB を開発した。本研究の成果は、公共 ChIP-seq データによるデータドリブンな生物学的知見や研究仮説の発見を大いに促すものである。

論文審査結果の要旨

本論文は増加の一途にある公共 ChIP-seq データに基づき、転写制御機構の解明に繋がる知見を提供しうる ChIP-seq データベースを構築する方法について示したものであり、全編６部からなる。

第１部は緒論であり、本研究の背景、目的および構成を述べている。

第２部では、公共 ChIP-seq データの解析に用いる解析パイプラインの概要を述べると共に、既存の ChIP-seq データベースに対して網羅性・信頼性に関する問題点を指摘し、これらの問題を解決するための手法について述べている。自動処理を実施する上での課題であった、ChIP-seq 実験に対する不完全な付帯情報の判別手法を提案し、多くの公共データを解析の俎上に乗せることに成功している。また、デコイ配列の ChIP-seq 解析への応用はこれまでに例がなく、難読領域における偽陽性の抑制に貢献することを明らかにした。

第３部では、信頼性の向上の一環として Strand cross-correlation の理論的特性評価を行っている。この手法に基づいた品質管理の指標は理論的な観点での裏付けがなく、指標として曖昧さが残っていた。本研究は Strand cross-correlation の一種である NCC（Naïve Cross-Correlation）に対して期待値を導出し、理論的洞察を与えた初めての報告であり、シミュレーションデータと実際のデータを用いた分析により本研究で見いだされた関係式がよく成り立つことを確認している。

第４部では第３部で得られた成果を用いて、品質管理のための新規指標である VSN を提案している。また、提案指標と Strand cross-correlation に基づく既存指標の比較を行っている。SN 比の評価手法として現在の主流である FRiP（Fraction Reads in Peaks）との比較では、既存指標よりも総合的な相関がより高いことを示している。更に、FRiP の弱点であり Strand cross-correlation に基づく指標の特徴として知られていた取得リード数に対するロバスト性についても、既存指標と同等のロバスト性を有していることを示し、VSN の実用性を確認している。

第５部では、これらの成果を踏まえて解析パイプラインをコンテナ仮想化とクラウドコンピューティング環境へ展開した成果と、得られた解析結果を提供する ChIP-seq データベースとして実装した Web アプリケーションである C4S DB について述べている。また、実際の解析結果を用いたデモンストレーションを通じて、本データベースの有用性と課題について言及している。

第６部は総括であり、本研究の成果をまとめると共に残された課題や今後の展望について述べている。

以上要するに本論文は、生命情報科学分野が蓄積している膨大なデータに対し、情報科学の観点から生物学的知見を提供するための一連の手法の提案と実装を行っており、応用情報科学ならびに生命情報科学の発展に寄与するところが少なくない。

よって、本論文は博士（情報科学）の学位論文として合格と認める。